

Demonstrating Quality Outcomes in Learning and Teaching: Examining ‘Best practice’ in the Use of Criterion-referenced Assessment

Sara Hammer (hammers@usq.edu.au)

Faculty of Business, University of Southern Queensland, Australia

This article has been anonymously peer-reviewed and accepted for publication in the *International Journal of Pedagogies and Learning*, an international, peer-reviewed journal that focuses on issues and trends in pedagogies and learning in national and international contexts. ISSN 1833-4105.
© Copyright of articles is retained by authors. As this is an open access journal, articles are free to use, with proper attribution, in educational and other non-commercial settings.

Abstract

Australian universities are now required to meet a range of quality assurance indicators directly related to demonstrated excellence in learning and teaching, which are evaluated by the Commonwealth Government’s Learning and Teaching Performance Fund. Consequently, they must now develop ways to demonstrate attainment of these prescribed benchmarks. One pedagogical practice adopted by Australian universities, as a way of identifying and demonstrating stated learning outcomes, is the use of criterion-referenced assessment. Fundamentally, assessment criteria should do two things: first, they should clearly articulate the desired qualities or characteristics of students’ work that are relevant to the task being assessed; second, they should show the relationship between the stated learning objectives of a course and the type of assessment task being used. However, correctly applied, assessment criteria also potentially contribute to students’ learning by enabling them to develop a sense of judgement in relation to their own performance. They also potentially provide the means for articulating appropriate disciplinary standards at a course level. Yet the usefulness of assessment criteria is potentially undermined by issues such as vagueness, confusion over the relationship between criteria and standards, a lack of consensus over the interpretation of criteria within teaching teams and disciplines and the challenge of articulating desirable qualities for assessment tasks that require complex, higher order thinking. This paper addresses these issues by reviewing relevant higher education literature and proposing six principles of good practice for the use of criterion-referenced assessment.

Introduction

The increasing professional relevance of Australian universities has led to a sharper focus on the quality of student learning and graduate outcomes. In order to qualify for funding from the Commonwealth government’s Learning and Teaching Performance Fund, universities must demonstrate excellence in learning and teaching that is judged according to a range of indicators, including good teaching and generic skills. The widespread adoption of criterion-referenced assessment reflects a concern to articulate clearly the desirable qualities of student work that are relevant to the assessment task. Their use should also ideally enable academic staff to demonstrate that students have achieved stated learning outcomes, including those associated with generic skills, to an appropriate standard through the completion of course assessment tasks. However, a review of existing literature suggests that the usefulness of assessment criteria as a tool for quality assurance is potentially undermined by issues such as confusion about

the relationship between assessment criteria and disciplinary standards, an inability to articulate the desirable qualities of assessment tasks, a lack of consensus among teaching and program team members over the interpretation of criteria and the challenge for academic staff of articulating criteria for complex assessment tasks that require higher order cognitive skills. This paper argues that there are six principles of good practice, which will enable academic staff to use criterion-referenced assessment for assuring student learning and developmental standards. It also briefly outlines some implications of these principles for more generalist programs, such as business and arts, where degree structures are more flexible and where students potentially enter at different points of the degree programs.

Pitfalls of Criterion-referenced Assessment

Sadler (2005, p. 178) defines a criterion as “A distinguishing property or characteristic of any thing, by which its quality can be judged or estimated, or by which a decision or classification may be made”. According to Dunn, Parry and Morgan (2002), the popularity of criterion-referenced marking in universities is a result of increasing concern for fairness, transparency and accountability in assessment regimes.

Such concerns have been driven in part by the need for universities to demonstrate excellence in particular graduate indicators, such as generic skills (Department of Education, Science and Training, 2005). However, they also stem from the current mass higher education context, where universities increasingly teach “students with little prior exposure to the unwritten rules and conventions of higher education” (James, McInnis & Devlin, 2002, p. 6). Because of its educative potential, criterion-referenced marking goes beyond a quality assurance function. Burton and Cuffe (2005) state that criterion-referenced marking empowers students to improve their own performance by providing explicit and attainable standards. However, there are possible complications and pitfalls for staff in both the implementation and the use of assessment criteria.

According to Harris and James (2006, p. 24), developing staff competency in the use of assessment criteria is complicated by the fact that assessment in higher education is used for many purposes: first, it is used as a means of grading and ranking for external stakeholders, such as professional bodies; second, it is used to provide feedback for students in a way that enables them to improve their performance; third, it is used as a means to focus students on appropriate areas of study; and finally, it provides staff with a means of assessing the effectiveness of both teaching and programs (Harris & James, 2006, p. 24). For staff in disciplines that are required to meet external accreditation requirements, the ability to distinguish clearly between criteria and standards is critical, yet confusion occurs over this process of differentiation.

While a criterion can refer to a desirable defining characteristic or quality of student work (for example, a well-structured essay), a standard normally refers to the performance ‘minimum’ or ‘threshold’ that a student must attain in order to achieve a passing grade (Dunn, Parry & Morgan, 2002; Sadler, 2005, p. 181). As O’Donovan, Price and Rust (2000, p. 75) discovered in practice, on its own the identification of desirable qualities in student work cannot pinpoint an appropriate standard for a given course at a given year of a particular disciplinary program. Indeed, qualities such as

“well-structured” can be used as valid criteria for both undergraduate and postgraduate years.

One way of engaging with standards is through the use of educational taxonomies, but without the addition of disciplinary context they lack precision. It is common practice within higher education to use taxonomies, such as those by Bloom, Krathwohl and Masia (1964) or Biggs (2003), as a way of thinking about the relationship between assessment and cognitive development. This constructivist view of assessment, that we construct knowledge based on our experience, is concerned with both where students are in terms of their cognitive development and where they should be at the end of the assessment process. According to such taxonomies, cognitive capacities develop from simple to complex (higher order) abilities or skills (Bloom, Krathwohl & Masia, 1964). According to Bloom and his colleagues (1964), taxonomies allow the classification of cognitive (and other) learning objectives, which in this case are cumulative, with lower order cognitive skills such as the recall of facts being subsumed by comprehension, extrapolation, application and so on. Taxonomies are useful for sharpening our thinking around course learning objectives, but it is academic staff, as experts in their discipline, who must prescribe appropriate levels of learning for particular stages of students’ degree programs. This is because what counts as the demonstration of cognitive abilities, such as critical thinking, differs according to discipline; to demonstrate critical thinking business students are often required to solve particular problems, whereas arts students might be required to produce an in-depth exploration of the problem itself. Such complexities make the task of articulating appropriate standards for different points in the degree structure a challenging one.

Recent findings (Morgan, Watson, Roberts, McKenzie & Cochrane, 2004) show that, while there are reasonably conceptualised differences between undergraduate and postgraduate studies, there is little differentiation within undergraduate programs. There are published examples of programmatic, pedagogical approaches to determining appropriate levels of student learning (see Kift & Nelson, 2005). However these are still exceptions in the Australian context, and even from a global perspective there is little evidence of published research on the issue of designing university courses so that they reflect standards that are appropriate for their program position (Morgan, Watson, Roberts, McKenzie & Cochrane, 2002). Despite this lack of scholarship, Morgan and his colleagues’ study (2004, p. 294) found that “Approximately half the universities surveyed volunteered the view that their students could be assisted more if their institution adopted a well-defined policy with clear definitions detailing the differences between the levels of undergraduate units”.

The suggestion by Morgan and his colleagues (2004) for addressing this gap in scholarship and practice is through a particular application of Bloom’s taxonomy which acknowledges the “disciplinary dimension” of learning in higher education. They define appropriate disciplinary levels of learning for undergraduate years accordingly:

- Level 1: Students can describe basic models that relate to a discipline and employ tools to apply the [conceptual] models to certain manageable contexts;
- Level 2: Students can analyse the assumptions underlying models, assess the worth of tools and apply models and tools to a range of more challenging contexts;

- Level 3: Students can evaluate the appropriateness of existing models and tools and design (adapt/customise) more appropriate ones as necessary. Students can also demonstrate how context-specific factors influence the design of models and tools. (Morgan, Watson, Roberts, McKenzie & Cochrane, 2002, para. 4)

If Bloom's taxonomy is mapped onto this conception of students' stages of disciplinary mastery, the result is that students may still engage in higher order thinking, such as analysis or evaluation, at a 'micro level' (for specific assessment tasks) for first year. However, from a 'macro' disciplinary dimension, the general expectation would be for students to "engage with this introductory interrelationship of models, tools and context" (Morgan, Watson, Roberts, McKenzie & Cochrane, 2002, para. 5). From the perspective of assessment criteria, this may mean that, although we might expect student work to demonstrate the qualities of both comprehension and analysis in first year, the former would be weighted more heavily, thus flagging comprehension of foundational disciplinary knowledge as the 'macro' learning dimension. Another way that assessment criteria might reflect appropriate levels for different years is through the use of grade descriptors, which indicate simpler or more complex modes of engagement for each criterion. For example, in first year management, students may fulfil the criterion of critical analysis by simply acknowledging the existence of different scholarly positions on a debate or topic, while in first semester of second year they might be required to apply foundational management theories to real life cases.

Either way, it is clear that, because of its scope, defining appropriate disciplinary standards to inform the task-based criteria that staff use should be a whole-of-discipline activity, which includes professional bodies. For this reason it is important for staff and professional groups to come to some consensus around standards and criteria or desirable qualities of student work in their discipline. However, there is evidence (Tan & Prosser, 2004) that there is not always a shared understanding, within disciplines of the role of assessment criteria, standards and grade descriptors. This may be partly due to the fact that academic staff members do not generally discuss assessment, ways of assessing and the purposes of assessment. More seriously, however, different understandings about the role of assessment criteria, standards and grade descriptors may arise as a consequence of deep differences in understanding about the purpose of assessment. These deep differences may be further influenced by power relationships among different disciplinary stakeholders (Tan & Prosser, 2004, p. 279). Despite the inherent complexity of making disciplinary standards explicit, Harris and James (2006, p. 26) argue that in a mass tertiary education environment, characterised by flexibility of entry pathways and modes of study, "there is a case for a primarily student outcomes-based approach to defining and monitoring academic standards, one grounded in the assessment and reporting of student learning".

An absence of discussion about desirable criteria and standards of student performance at a disciplinary level may also result in the use of assessment criteria at an individual course level that are either too vague or too prescriptive. In their study based on student experience of criterion-referenced assessment, O'Donovan, Price and Rust (2000, p. 79) found that, if criteria were too vague and imprecise, students and staff could interpret them too widely; as expressions of lecturer expectations they

offered inadequate clarification and guidance. According to Tan and Prosser (2004, p. 268), one way of resolving the issue of vague assessment criteria is through the use of grade descriptors. They argue that grade descriptors offer an “alternative approach to achieving standards-based assessment” by describing the achievements necessary to obtain a particular standard or grade in relation to each criterion. Sadler (2005, p. 192) describes descriptors as “verbal descriptions consisting of statements setting down the properties that characterise something of the designated levels of quality”. For this reason, descriptors are useful for both markers and students in determining whether students have successfully met each set criterion to a particular standard.

Yet others (Dunn, Parry & Morgan, 2002; Elander, Harrington, Norton, Robinson & Reddy, 2006, p. 86; O'Donovan, Price & Rust, 2000; Sadler, 2005, p. 181) warn against criterion-referenced assessment that defines qualities of student achievement too prescriptively. One concern is that the use of assessment criteria is an attempt to impose precision on what are often subjective judgements (O'Donovan, Price & Rust, 2000, p. 79). Another related concern with the tendency towards increasing precision in defining assessment criteria is that students using them will as a result adopt a ‘strategic’, shallow learning approach where each strives to demonstrate the execution of discrete skills rather than viewing their learning holistically (Dunn, Parry & Morgan, 2002; Elander, Harrington, Norton, Robinson & Reddy, 2006, p. 86).

Rust, Price and O'Donovan (2003, p. 151) point out that, if the purpose of assessment criteria is to clarify desirable qualities of a complex assessment task, focusing only on what can be articulated explicitly may not constitute “useful knowledge” of the assessment process. For the transfer of such “tacit” knowledge, they suggest experiential processes, which include “observation, imitation, dialogue and practice” (p. 152). Ultimately, then, they argue that the use of criterion-referenced assessment systems must be complemented by the use of “exemplars, marking practice and the opportunity for dialogue between staff and students, to complement explicit knowledge provided through the verbal explication of assessment criteria by staff and in written format embodied within the grid” (p.161). Bringing together the above complexities and concerns, it is possible to make some preliminary observations about what constitutes good practice in criterion-referenced assessment.

Six Principles of ‘Good Practice’ for Criterion-referenced Assessment

First, assessment criteria should be aligned with stated course outcomes and the type and (arguably) the relative weighting of course assessment. According to the constructivist perspective, it is what teachers ‘do’ and what they assess, rather than what they say, which determines students’ ability to meet desired learning outcomes (Biggs, 2003). Therefore assessment criteria should clearly connect with both stated course outcomes and the type of assessment task being set. As a related issue, assessment criteria should also reflect the desired quality and standards of graduate attributes and skills being assessed in each instance. Students’ successful completion of set assessment thus enables universities to demonstrate excellence on a key indicator used by the Learning and Teaching Performance Fund (Department of Education, Science and Training, 2005).

Second, assessment criteria should be specific to the task being assessed. Overly generic or misaligned assessment criteria do little to communicate expected performance qualities and standards to students. Criteria used should demonstrate a clear link to the task being assessed. Nonetheless, it is worth noting that similar tasks would logically be assessed using similar assessment criteria. If an essay is a structured argument supported by research/evidence, the qualities of a good essay will be similar across disciplines. For example, in either economics or management, a good essay is one that demonstrates the qualities of good structure and coherence. This principle of good practice in the use of assessment criterion can ensure both transparency and consistency in assessment practices across disciplines and programs.

Third, assessment criteria should be supplemented by the use of exemplars, practice and dialogue between staff and students. This not only addresses the issue of communicating standards and norms that are difficult to express in words but also helps academic staff who are not habituated to making such standards and norms explicit.

Much of what counts as professional judgement in the context of higher education remains tacit. Added to this, the complexity of many tasks that we set students may simply not lend themselves to written explanation. Supplementing assessment criteria with exemplars and learning activities makes our assessment practices more transparent, but it also develops students' capacity to make judgements about their own and others' performance (Harris & James, 2006).

Fourth, assessment criteria should be used in combination with grade descriptors. It is grade descriptors (Tan & Prosser, 2004), which tell both students and assessors in detail what levels of performance they must demonstrate for each criterion to meet the minimum standard (pass) or any other grade above or below it. However, grade descriptors also offer a means for teaching staff to express specific course- and discipline-based interpretations of particular assessment criteria. Thus, for students to demonstrate a passing grade for the criterion of comprehension for a first year economics course, they may be required to apply a general economic theory. For the same students to demonstrate a passing grade for the criterion of comprehension in a second year management course, they might be required to demonstrate an understanding of more complex concepts in relevant management theory. Despite the potential for grade descriptors to make assessment criteria clearer for students and staff, there will always be the potential for conflict and confusion over the standards that such descriptors represent, and even wide interpretation of the descriptors themselves. Nonetheless, the use of grade descriptors provides a means for university staff to make grading systems and appropriate learning standards more transparent to students and other stakeholders, such as government.

Fifth, assessment criteria should be reinforced and reviewed through moderation processes, which build shared understandings among teaching team members. Consequently, the use of assessment criteria should be supported by policies and management practices that enable teaching teams to develop shared meanings around the weighting and interpretation of chosen criteria and standards for course assessment. The development of shared understandings around appropriate assessment standards cannot be left to individual course convenors. As Tan and Prosser (2004) point out, even within disciplines there is not always a shared

understanding about the role of assessment criteria, standards and grade descriptors. Reaching a clear consensus about qualities and standards of student assessment within disciplines provides universities with greater capacity to demonstrate consistency in student learning outcomes, including generic skills.

This leads to the sixth principle: that assessment criteria should articulate, at a course level, wider disciplinary and professional consensus around appropriate standards of student performance for each stage of their degree program. Whether through the use of learning taxonomies, consultation with external stakeholders or consensus of disciplinary program teams, assessment criteria at an individual course level must relate to wider, agreed upon disciplinary or program standards; these should account for the horizontal position of the course (relative to other courses of its year), as well as the vertical position of the course (relative to related courses in subsequent years). As Kift and Nelson (2005, p. 231) affirm, “[I]ndividual units should be designed to produce cumulative benefits in terms of skills or graduate capabilities development”. However, this approach has significant implications for degree programs, such as arts and business, which have both a flexible structure and multiple disciplines.

Potential Problems for Flexible, Multidisciplinary Programs

Programs with a flexible degree structure, which make less use of pre-requisites than occurs in highly structured programs such as engineering, face challenges in assuring the quality of student learning through the use of assessment criteria. This is because, while in this context students can be encouraged to take courses in the appropriate order, there are a significant number who do not, a trend that is compounded by the general tendency for universities to rationalise the overall number of offerings that students can choose from. Program teams who are serious about formulating and articulating appropriate standards for different years in students’ undergraduate degrees (in particular) then face the problem of how to support students who take courses out of sequence and do not necessarily possess the cognitive skills to complete them successfully. In this case, a clearer articulation of qualities and standards of student work may actually serve to increase pressure on both course teaching teams and newer students, resulting in higher failure levels and lower retention and progression rates.

The multidisciplinary dimension of such programs adds an additional complication to assuring the quality of student learning. Business and arts degrees often have core foundation courses that cover the range of disciplines on offer, but such courses are not always available for students entering at different points of the academic year. Instead, many choose second or third year courses which, if designed to develop students’ learning at the appropriate standard for that year, will generally assume prior knowledge of disciplinary norms and content that students do not yet possess. Based on a taxonomic conception of student learning (Morgan, Watson, Roberts, McKenzie & Cochrane, 2002), students must develop comprehension of foundational content before they can successfully engage in higher order thinking within that discipline. More research is needed to evaluate possible strategies for dealing with these issues.

Conclusion

Increasingly, universities must be able to demonstrate that when students graduate they will have the knowledge and capabilities to succeed as knowledge workers and

professionals in current and future work environments. This paper has explored the use of criterion-referenced assessment as a tool for conceptualising and assessing precise learning objectives, and as one way of thinking about appropriate levels of learning for different stages of students' degrees. However, unsupported assessment criteria can define only broad qualities, such as clarity of expression. They cannot articulate appropriate thresholds or standards of student learning at different levels of their degree; their use must be supplemented with grade descriptors, models and exemplars. Nor can assessment criteria serve as a proxy for wider disciplinary consensus about the quality and standards of student work; this must be facilitated through management processes and policies at a disciplinary and faculty level. Even if faculties enthusiastically embrace the six principles of good practice outlined in this paper, there are still challenges. These are particularly acute for faculties with flexible, multidisciplinary offerings. They relate to the difficulty of balancing the adoption of a cumulative, whole-of-program approach which ensures students' cognitive development across their degree with the reality of students who will not, or cannot, pursue their particular program according to the recommended sequence. This presents an interesting challenge for higher education practitioners in Australia and elsewhere.

References

- Biggs, J. (2003). *Teaching for quality learning at university* (2nd ed.). Maidenhead, UK: The Society for Research into Higher Education & Open University Press.
- Bloom, B. S., Krathwohl, D. R., & Masia, B. B. (1964). *Taxonomy of educational objectives* (Book 2). New York: David McKay Company.
- Burton, K. J., & Cuffe, N. (2005). The design and implementation of criterion-referenced assessment in a first year undergraduate core law unit. *Legal Education Review*, 1(59). Retrieved March 1, 2007, from <http://eprints.qut.edu.au/secure/0004202/01/KJB - CRA - LER article.doc>
- Department of Education, Science and Training. (2005, December). *Discussion paper – Learning and Teaching Performance Fund: Future directions*. Retrieved August 1, 2007, from http://www.dest.gov.au/NR/rdonlyres/645A5662-D277-49DC-A59A-937022D5A0F1/10279/LTPF_DiscussionPaper_2006.pdf
- Dunn, L., Parry, S., & Morgan, C. (2002, August). *Seeking quality in criterion-referenced assessment*. Paper presented at the "Learning communities and assessment cultures" conference organised by the European Association for Research on Learning and Instruction Special Interest Group on Assessment and Evaluation, University of Northumbria, UK.
- Elander, J., Harrington, K., Norton, L., Robinson, H., & Reddy, P. (2006). Complex skills and academic writing: A review of evidence about the types of learning required to meet core assessment criteria. *Assessment & Evaluation in Higher Education*, 31(1), 71-90.
- Harris, K., & James, R. (2006, October). Facilitating reflection on assessment policies and practices: A planning framework for educative review of assessment. *Studies in Learning, Evaluation, Innovation and Development*, 3(2), 23-36.
- James, R., McInnis, C., & Devlin, M. (2002). *Assessing learning in Australian universities: Ideas, strategies and resources for quality in student assessment*. Melbourne, Vic: Centre for the Study of Higher Education, University of Melbourne & the Australian Universities Teaching Committee. Retrieved January 31, 2006, from <http://www.cshe.unimelb.edu.au/assessinglearning>

- Kift, S., & Nelson, K. (2005, July). Beyond curriculum reform: Embedding the transition experience. In A. Brew & C. Asmar (Eds.), *Higher education in a changing world: Proceedings of the 2005 annual international conference of the Higher Education Research and Development Society of Australasia (HERDSA)* (pp. 225-235). Milperra, NSW: Higher Education Research and Development Society of Australasia.
- Morgan, C. K., Watson, G. K., Roberts, D. W., McKenzie, A. D., & Cochrane, K. W. (2002, December). *Structuring the curriculum for different years of undergraduate programs*. Paper presented at the annual conference of the Australian Association for Research in Education, Brisbane, Qld. Retrieved January 31, 2007, from <http://www.aare.edu.au/02pap/mor02051.htm>
- Morgan, C. K., Watson, G. K., Roberts, D. W., McKenzie, A. D., & Cochrane, K. W. (2004). Scholarship neglected? How levels are assigned for units of study in Australian undergraduate courses. *Assessment & Evaluation in Higher Education*, 29(3), 284-298.
- O'Donovan, B., Price, M., & Rust, C. (2000). The student experience of criterion-referenced assessment through the introduction of a common criteria assessment grid. *Innovations in Education and Teaching International*, 38(1), 74-85.
- Rust, C., Price, M., & O'Donovan, B. (2003). Improving students' learning by developing their understanding of assessment criteria processes, *Assessment & Evaluation in Higher Education*, 28(2), 147-164.
- Sadler, R. (2005). Interpretations of criteria-based assessment and grading in higher education, *Assessment & Evaluation in Higher Education*, 20(2), 175-194.
- Tan, H. K., & Prosser, M. (2004). Qualitatively different ways of differentiating student achievement: A phenomenographic study of academics' conceptions of grade descriptors. *Assessment & Evaluation in Higher Education*, 29(3), 267-282.